www.elsevier.com/locate/developmentalbiology

Developmental Effector Gene Regulation: Multiplexed Strategies for Functional Analysis

Lijun Wang, Kari Koppitch, Ann Cutting, Ping Dong, Parul Kudtarkar, Jennie Zheng, R. Andrew Cameron, Eric H. Davidson

PII: S0012-1606(18)30208-2
DOI: https://doi.org/10.1016/j.ydbio.2018.10.018
Reference: YDBIO7889

To appear in: *Developmental Biology*

Received date: 22 March 2018
Revised date: 23 October 2018
Accepted date: 24 October 2018

Cite this article as: Lijun Wang, Kari Koppitch, Ann Cutting, Ping Dong, Parul Kudtarkar, Jennie Zheng, R. Andrew Cameron and Eric H. Davidson, Developmental Effector Gene Regulation: Multiplexed Strategies for Functional Analysis, *Developmental Biology,* https://doi.org/10.1016/j.ydbio.2018.10.018

Developmental Effector Gene Regulation: Multiplexed Strategies for Functional Analysis

Lijun Wang, Kari Koppitch, Ann Cutting, Ping Dong, Parul Kudtarkar, Jennie Zheng, R. Andrew Cameron*, Eric H. Davidson[1]*

Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125

*Corresponding author. acameron@caltech.edu

ABSTRACT

The staggering complexity of the genome controls for developmental processes is revealed through massively parallel cis-regulatory analysis using new methods of perturbation and readout. The choice of combinations of these new methods is tailored to the system, question and resources at hand. Our focus is on issues that include the necessity or sufficiency of given cis-regulatory modules, cis-regulatory function in the normal spatial genomic context, and easily accessible high throughput and multiplexed analysis methods. In the sea urchin embryonic model, recombineered BACs offer new opportunities for consecutive modes of cis-regulatory analyses that answer these requirements, as we here demonstrate on a diverse suite of previously unstudied sea urchin effector genes expressed in skeletogenic cells. Positively active cis-regulatory modules were located in single Nanostring experiments per BAC containing the gene of interest, by application of our previously reported “barcode” tag vectors of which >100 can be analyzed at one time. Computational analysis of DNA sequences that drive expression, based on the known skeletogenic regulatory state, then permitted effective identification of functional target site clusters. Deletion of these sub-regions from the parent BACs revealed module necessity, as simultaneous tests of the same regions in short constructs revealed sufficiency. Predicted functional inputs were then confirmed by site mutations, all generated and tested in multiplex formats. There emerged the simple conclusion that each effector gene utilizes a small subset of inputs from the skeletogenic GRN. These inputs may function to only adjust expression levels or in some cases necessary for expression. Since we know the GRN architecture upstream of the effector genes, we could then conceptually isolate and compare the wiring of the effector gene driver sub-circuits and identify the inputs whose removal abolish expression.

Keywords: Skeletogenic effector genes/embryonic gene regulation/recombineered BACs/tag vectors

INTRODUCTION

Developmental cis-regulatory analysis relies on a suite of techniques that includes methods to define specific gene expression spatiotemporally, to perturb that expression and to mutate transcription factor binding sites in reporter constructs. Genome-wide strategies such as, for example, Perturb-seq (Dixit et al., 2016) and Mosaic-seq (Xie et al. 2017) deeply automate the first steps of this pipeline. But distinguishing among different molecular mechanisms that yield similar phenotypes requires time- and labor-intensive follow-up. It is interesting to reflect that virtually all of our hard scientific knowledge of how genomic regulatory sequence encodes developmental function, particularly spatial gene expression, depends on

[1] deceased

variants of a single experimental paradigm. Putative regulatory DNA sequences (several hundred base pairs long) are ligated to an immediately adjacent sequence containing a weakly promiscuous transcriptional promoter followed by a sequence encoding a reporter protein and poly(A) tail, and introduced by gene transfer into eggs, embryos, or differentiating cells in culture. The expression of the reporter provides a qualitative and quantitative metric of regulatory sequence function which can be experimentally perturbed down to the nucleotide level to establish function (for illustration of the informational richness these methods can reveal about local cis- regulatory structure and function, and comprehensive reviews see: Davidson, 2006; Peter and Davidson, 2015; Swanson, et al., 2011; Spitz and Furlong, 2012). The experimental paradigm dates from the late 1980's and, over the last 20 years, has been heavily exploited in developing flies, mice, sea urchin embryos, nematodes, frogs, chicks, etc. But all things pass, and there have now arisen specific additional requirements in cis-regulatory analysis for which the classical paradigms do not easily suffice. These problems are exacerbated in examination of the regulatory connections between effector genes (genes encoding cell biology and differentiation proteins) and upstream GRNs (gene regulatory networks) that control developmental regulatory state. The encoded linkages immediately determining effector gene expression constitute a little-known but essential domain of developmental control circuitry. The most basic pressure on traditional means of cis-regulatory analysis in this context lies in the multiplicity of effector genes, which is ill-accommodated by the slow and laborious, gene by gene methodologies of classical cis-regulatory analysis. But there are other problems as well: (i) constructs in which normally distant cis-regulatory modules are placed adjacent to the promoter ("short constructs") often display excess functionalities compared to those functions they mediate in their natural contexts (e.g., (Wahl et al., 2009); see examples cited in (Peter and Davidson, 2015)); (ii) short constructs can be used to demonstrate sufficiency in regulation of a given phase of expression, but not to demonstrate their necessity, whereas in life, additional cis-regulatory modules often display overlapping regulatory activities; (iii) the mechanisms of module choice are not directly accessible out of the normal genomic spatial context; (iv) nor are the consequences of interaction with negatively acting modules, if any; (v) in traditional modes of cis-regulatory analysis, the initial project of scanning the vast genomic distances occupied by introns and intergenic "space" for active cis-regulatory modules can be prohibitively laborious. Here we have employed sea urchin embryos as our test bed, as in this genome, which is about 1/4th the size of a mammalian genome, cis-regulatory modules are usually to be found within given BAC constructs containing the gene of interest (~140 kb, typically containing 3-5 genes). Thus, by beginning with a recombineered BAC that properly spatially and temporally expresses a reporter gene situated immediately after the start of translation of the gene of interest, we know a priori that all the necessary genomic regulatory machinery is located within that BAC.

Among the most prominent developmental features of sea urchin embryogenesis is the early specification and differentiation of a dedicated cell lineage that uniquely and exclusively expresses a skeletogenic fate. All descendants of this lineage ingress into the blastocoel before gastrulation and thereafter, following spatial cues expressed on the inner ectoderm wall, they form a bilaterally disposed syncytial cable within

which a species-specific pattern of calcite biomineral rods are deposited. The genomically encoded circuitry accounting for initial spatial specification of the skeletogenic regulatory state, and the GRN encoding the following regulatory functions that initiate skeletogenic differentiation, as attested by the precocious expression of skeletogenic effector genes, are extraordinarily well known (Oliveri et al., 2008). This skeletogenic GRN has been computationally demonstrated to suffice for predictive explanation of skeletogenic regulatory gene expression (as for the remainder of embryonic mesoderm and also endoderm specification as well (Peter et al., 2012). In addition, cis-regulatory studies were carried out earlier on several skeletogenic effector genes (Kurokawa, et al.,1999; Sucov, et al., 1988; Amore and Davidson, 2006; George et al., 1991; Katoh-Fukui, et al., 1991), and using knowledge of the skeletogenic GRN (Oliveri et al., 2008), perturbation analysis has been used to obtain a broad scale inference of the regulatory inputs to a large number of different skeletogenic effector genes (Rafiq et al., 2012). The immediate antecedents to this work were application of a new technology for isolation of multiple of cell and tissue types from disaggregated sea urchin embryos by FACS, on the basis of expression of recombineered BACs expressing regulatory genes (and fluorogenic reporters) specific to given regulatory states, here the skeletogenic regulatory state (Barsi et al., 2014). This work resulted in the isolation of scores of skeletogenesis specific effector genes, including many known earlier (Rafiq et al., 2012; Zhu, et al., 2001; Livingston, et al., 2006) but also many identified de novo. The isolated skeletogenic effector genes were authenticated by in situ hybridization and annotated (Barsi et al., 2014). A set of eight genes chosen for the present study were selected arbitrarily from this earlier analysis (Table 1). The only criteria applied were: i) a reasonable level of expression; ii) the absence of previously known cis-regulatory information; and iii) diversity in function, so that cell biology as well as biomineralization genes were included. Our overall objective here was to determine for this whole set of genes the regulatory inputs linking their skeletogenic function with the upstream specification GRN, and hence to unravel the near terminal control wiring that tissue-specifically animates these effector genes.

MATERIALS AND METHODS

Animals, culture and injections

Purple sea urchins (*Strongylocentrotus purpuratus*) were collected at various locations on the southern California coast and shipped to the Caltech campus. Animals where maintained in chilled saltwater tanks until use. Gametes were obtained by vigorous shaking, electric shock or intra-coelomic injection of 0.55 M KCl. Fertilization and microinjection of zygotes was performed as described (Cameron et al., 2004). When using DNA fragments less than 70,000 bp, a 5-fold mass excess of *Hind*III digested genomic DNA (average length 5 -10 kb) was also injected. All DNA injections, whether individual or pooled, were calculated to yield a specific copy number based upon known length mass as measured by spectrophotometer and an assumed injection volume of 2 pl.

Bacterial artificial chromosome library and clone selection.

Bacterial artificial chromosomes (BACs) were developed to provide a stable, easily manipulated vector system to carry large fragments of genomic DNA. These BAC vectors overcame the fragility and size limitations of a variety of cosmid counterparts. The Sea Urchin Genome Project used the pBACe3.6 vector (Frengen et al., 1999) and produced a suite of BAC libraries, each of which was constructed from the same DNA as that sequenced for the genome assembly (Cameron et al., 2000). These libraries are by now, well characterized and hundreds of clones have been described in the literature and published on the echinoderm genomic information website, www.echinobase.org.

As part of the purple sea urchin genome sequencing effort, about 8000 of the BAC clones were assigned to a minimum tiling path and sequenced in pools (Sodergren et al., 2006). Since these clones came from the same DNA as the sequencing project, the BAC clones could be easily mapped to the genome assembly. About 7700 BACs are successfully mapped to the assembled genome sequence (Cameron et al., 2009). Using the functionality of the database search engine, it is thus possible to directly identify a BAC clone mapped to a feature such as a gene.

BAC recombineering

Candidate BAC clones for each gene were engineered to contain a sequence encoding GFP by methods adapted from published procedures using a λ-red recombineering system (Sharan et al., 2009; Holmes et al., 2015). In this system, a replication defective λ phage (λTetR) was introduced to the specific BAC-containing cells to provide heat inducible recombination functions. *E. coli* DH10B was used as the BAC host strain. To make the deletions, a donor DNA cassette containing both short (50 bp) sequences complementary to the region flanking the desired region and a spectinomycin selective maker was electroporated into the BAC host cells. The correct recombinants were selected by colony PCR using construct specific primers. The integrity of the BAC constructs was confirmed by sequencing specific regions.

Isolation of sequences containing positively acting cis-regulatory modules.

Recombinant BAC clones that expressed GFP exclusively in skeletogenic cells were analyzed for gene structure. The gene region was defined as 30 kb upstream from the transcription start site and downstream from the 3′ end of the final exon unless another gene region was reached in either direction. Overlapping 2 kb fragments, exclusive of coding sequence and the region 50 bp 5′ of the start of transcription, were selected from within the gene region. PCR primers based on the reference genome DNA sequences were designed to recover the selected fragments. The PCR primers included complementary sequences to one of the 129 unique ‘nanotag’ markers (Nam and Davidson, 2012) After purification each fragment was concatenated to a unique tag by fusion PCR (Hobert, 2002). Resultant tagged products were purified and pooled for injection. When a fragment as found to be active, that region was deleted from the GFP-recombineered BAC clone and the resultant construct was tested by gene transfer into embryos. Positive controls for experiments with these deleted BAC constructs are the active fragment construct itself and the

Tbr-GFP construct (Wahl et al,.2009). The confirmation was considered successful if the construct failed to express GFP in any of the embryos as observed under the microscope (Tu et al., 2014; Nam and Davidson, 2012; Cameron, et al., 2009).

Zygotic injection

Genomic fragments and sub regions derived from them that were identified as positive in the experiments above were tested by injection into embryos as pools or individually. Pools of constructs were aliquoted to inject each construct at ~20 copies/embryo. All pools contained control fragments that exhibit positive and negative expression based on prior results. Positive control fragments often included the TBr-GFP-BAC construct and the complete, GFP-recombineered BAC clone. Injected embryos (100 – 150 per pool) were harvested and lysed at 24 hr after fertilization. Fragments that induced GFP expression in PMCs in over 12% of the embryos were considered positive.

For spatial expression analysis, GFP recombineered BAC clones were injected at 500 copies/embryo and individual reporter constructs at 1500 copies/embryo. Microscopic observations were made at 24-48 hours after injection. The results fell into several categories: spatially specific, spatially unrestricted or ectopic and too low to assess.

Nanostring NCounter analysis

Total RNA and genomic DNA were simultaneously isolated from transgenic embryos using a Qiagen AllPrep Kit. RNA was checked for a 260/280 nm absorbance ratio over 1.8 and then reverse transcribed using an iScript cDNA Synthesis Kit (Bio-rad). cDNA was ethanol precipitated at 4°C using glycogen as a carrier. A 359-370 bp target region containing the unique nanotag sequence and part of the GFP was amplified from both the cDNA and gDNA samples. 100 ng templates of these samples were then converted back to RNA by *in vitro* transcription using a T7 RNA Polymerase-Plus Enzyme and NTP mixes (Ambion). Gene expression was quantitated using the NanoString nCounter with a specific codeset developed for the 129 tag sequences (Nam and Davidson, 2012).

Quantitative PCR measurement

At 24 hours after fertilization 100-300 embryos were harvested and processed as above up until ethanol precipitation of the cDNA. qPCR was run on the genomic DNA (5-50 ng per reaction) and cDNA (50-100 ng/reaction) using tag-specific or GFP primers (Supplementary Table S2). In addition, unprocessed single stranded RNA was measured to check for genomic DNA contamination of the cDNA sample. Each condition was run in triplicate using an Intercalating Dye (SYBR Green) protocol and the average CT value taken. The copy number of injected DNA amplified during development was estimated in comparison to a single copy gene, Sp-Foxa. Estimated number of transcripts of injected DNA was calculated by comparison to known expression level of Sp-Ubiquitin cDNA. Transcripts per construct incorporated were then calculated. Expression was considered to be lost if the transcripts per construct copy number was at least 3X less than positive controls and near the level of the negative control.

Functional analysis of putative target sites by mutation.

We derived transcription factor binding motifs based on CIS-BP database (Weirauch et al., 2014). The motif sequences are listed in Table 2. Single or combined motif mutations were designed. Multiple different combinations of mutations were created and illustrated in Snapgene files. We also generated cluster mutations in which all factors corresponding to a high ATACseq peak region were mutated. In order to achieve a cluster of multiple mutations, we utilized synthetic DNA fragments (gBlocks Gene Fragments; IDT). Basically, we made motif mutations by switching sequence AT to CG without generating a new binding site. Mutation constructs were made by connecting PCR fragments that cover non-mutated regions with gblock that contain mutations and a nanotag that encode a GFP and a tag sequence from a 13 tag system (Nam and Davidson, 2012). The connection of multiple fragments was achieved by using In-Fusion method (Clontech). This method enables directional, seamless assembling of multiple overlapping DNA molecules into a linearized vector by the concerted action of a 5' exonuclease, a DNA polymerase and a DNA ligase. The In-Fusion method is highly efficient (>95%) in our hand, provided high accuracy of the sequence of our genes being studied.

BioTapestry

Gene regulatory networks derived from these studies were drawn using a local copy of the BioTapestry Editor Version 7.1.0 (Longabaugh, 2013; Paquette et al., 2016). The program is available for download from the BioTapestry web site (http://www.biotapestry.org/).

Results and Discussion

Methodological theory and sequence of functional experiments.

There were two major reasons to revise cis-regulatory analysis methodology: first, to sharply increase throughput for research on multiple genes; and second, to obtain direct evidence as to necessity as well as sufficiency of given positively acting cis-regulatory modules in the natural genomic spatial context. We arrived at the strategy summarized in the annotated flowchart of Fig. 1, in which new applications of BAC recombineering are combined with new applications of the Nanostring barcode tag technology for cis-regulatory analysis invented earlier in this laboratory (Nam and Davidson, 2012). This approach takes advantage of the sea urchin genome organization, in which a single BAC generally suffices to contain the entire genomic interval of a gene of interest (Buckley et al, 2017) from the upstream to the downstream flanking gene, and the well-established high fidelity of expression of transgenes introduced into sea urchin eggs. In Fig. 1 “S1” indicates a step in which construct expression is monitored microscopically both spatially and quantitatively (as the fraction of expressing embryos) using fluorochrome markers, and “Q” indicates a step in which cis-regulatory activity is measured quantitatively per batch of embryos using NanoString technology (an instrumental method of simultaneously quantifying large numbers [here, 129] of transcripts which are detected by laser detection of color-coded probes for each molecular species; Geiss et al. 2008). In the following paragraph, the objective, design and rationale of each step in Figure 1 are briefly discussed.

S1.Efficacy of recombineered marker BACs. BACs containing skeletogenic effector genes identified in our previous study (Barsi et al., 2014) (with the exception of a previously unstudied regulatory gene, Foxb (Oliveri et al., 2008; Tu et al., 2006) were isolated, and sequences encoding GFP were recombined in, immediately following the translation start sites ("marker BACs"). Throughout this work we utilized published BAC recombineering methods based on a replication defective λ phage, which inductively expresses recombination proteins (Hollenback et al., 2011; Holmes et al., 2015). We found this methodology to be fast, efficient, and easily multiplexed so that diverse recombinants can be prepared simultaneously. When injected into fertilized sea urchin eggs, concatenated marker BACs containing skeletogenic genes will be expressed in skeletogenic cells in 50% of embryos if genomic incorporation occurs at 3rd cleavage, and in 25% if it occurs at 4th cleavage, when the antecedents of the skeletogenic cells are segregated. Such results demonstrate that (i) the correct BAC was chosen; (ii) a skeletogenic cis-regulatory system that works with the endogenous promoter lies somewhere within the BAC; (iii) the gene is expressed at sufficient levels to support further cis-regulatory analysis.

Q2.

Isolate sequences containing positively acting cis-regulatory modules. A primary objective of this work was to eliminate the slow and tedious process of locating active cis-regulatory modules within genomic space so that this information could be obtained for multiple BACs in single transgene injection experiments. As diagrammed, intronic and intergenic regions surrounding the gene of interest in the marker BAC were divided into overlapping tiling arrays of ~2 kb fragments and incorporated into tagged expression vectors by multiplex fusion PCR. The tag system consists of 129 vectors into which fragments of DNA can be incorporated, each containing a promoter, GFP gene, and a unique sequence tag identifiable by a NanoString probe (Nam and Davidson, 2012). A mixture of the whole set is injected into eggs, wherein the individual vectors function independently, and NanoString analysis reveals the level of activity quantitatively as the quantity of transcribed tag normalized by the genomically incorporated amount of that tag vector (Nam and Davidson, 2012). In this way, one or several 2 kb fragments were recovered from each BAC which contained active cis-regulatory modules ("hot pieces"), whereas the vast majority of fragments were inactive. We could further distinguish the efficacy of the fragments by individual injection and microscopic observation to assess spatial extent of expression.

S3:

Assess sufficiency by deleting the predicted cis-regulatory sequence from the marker BACs. The upstream GRN activates approximately 10 possible skeletogenic specific, positively active, regulatory drivers of early skeletogenic effector genes (Oliveri et al., 2008). Transcription factor binding sites recognized by these drivers were inferred using the CisBP database (Weirauch et al., 2014)(see Table 2) and mapped in the hot piece sequences. Where subsets of these putative sites appeared in clusters, deletions a few hundred base pairs in length were created within the hot piece regions of the original marker BACs. Embryonic frequency of skeletogenic expression was measured vs. the normal marker BAC controls. Perhaps

surprisingly, in more than half the genes tested, given deletions entirely abolished the expression of the marker BAC, indicating that we had identified a necessary as well as sufficient cis-regulatory module.

S4:

Assess necessity by deleting small regions within individual hot pieces to finely map functional sites. In contrast, some deletions failed to affect BAC activity or resulted in mere decreases in activity. This was always correlated with the initial observation of more than one hot piece per gene, indicating the presence of widely separated regulatory modules of at least partially overlapping function. In this case, effects of clustered site deletions were assayed on the individual hot pieces in isolation, to eliminate compensation from other regions within the whole BAC. To this end, the deleted BAC regions could be tested in short constructs by using that portion of the hot piece that wasn't deleted fused to a nanotag encoding a GFP reporter. The effect on spatial expression frequency was then compared to the control hot piece vectors.

Q5:

Functionally confirm regulatory inputs. The remaining step in this protocol is to assign functionality of specific regulatory factor binding sites identified in experiments S3 and/or S4 by mutating these sites and quantifying the subsequent effect on GFP expression. This process is facilitated by the identification of all sites for each a putative input (particular regulatory factor) within the functional regions identified by deletion in each hot piece. In the context of either the larger marker BAC constructs or smaller hot pieces, mutations were designed specifically according the numerical PWM data by altering either invariant or forbidden sites (Table 2). Since the spatial specificity of the hot pieces was already demonstrated, quantitative multiplex analysis of mutation sets in single NanoString experiments sufficed. The data generated by the strategies summarized in Fig.1 together afford high probability identification of functional cis-regulatory inputs, given *a priori* restriction to subsets of only ten factors afforded by prior knowledge of the GRN, the short sequence length of the deletion regions, and predictive application of the PWM data for these factors.

Recombinant marker BACs are used to localize expression of effector genes to the skeletogenic cells.

The genes analyzed in this study (Table 1) encode biomineralization proteins and cell biology proteins that contribute to the unique behavior of this cell lineage (Lyons et al., 2014), as well as transcription factors that mediate the downstream regions of the GRN and are activated later in specification (Oliveri et al., 2008). These genes are all specifically expressed in the skeletogenic cell lineage, as explicitly shown by quantitative comparison to the transcriptomes of other cell lineages and by *in situ* hybridization (Barsi et al., 2014).

The first step in our cis-regulatory analysis pathway (Fig.1, S1) is preparation and testing of recombineered marker BACs that each contain a target gene and extensive flanking sequence. The sea urchin genome is about four times more compact than the mammalian genome, and since the average intergenic distance is

about 30 kb (Sodergren, et al., 2006), a typical 140 kb BAC contains several genes flanking the gene of interest if the latter is more or less centrally located. The BACs in this study can be easily accessed by name and mapped to the genome along with other genomic features in the *S. purpuratus* genome database (www.echinobase.org). Recombinant marker BACs were constructed as above, and tested after injection into fertilized eggs by observation of the morphological localization of GFP fluorescence (see Fig.2). In embryos transgenic for each marker BAC, GFP is specifically expressed very clearly in the morphologically distinctive skeletogenic lineage. Note that, because the skeletogenic cells form syncytia within the developing blastula, the fluorescent protein is evident in all the cells of this lineage despite mosaic incorporation of the transgene. These data are in accord with our experience that, in this genome, a BAC containing a gene of interest is very likely also to encompass the cis-regulatory module(s) that control its spatial expression.

Overlapping, short fragments are used to locate active cis-regulatory sequences within reporter BACs.
The marker BACs for each of the eight skeletogenic genes of interest serve as an experimental foundation for generating the tagged DNA fragments used in next step of the protocol (Fig.1, Q2). Using the tiling arrangement described above, we generated reporter vectors in which ~2 kb DNA fragments directed the expression of GFP. Each fragment is barcoded with a unique sequence tag from the 129-tag NanoString system. (Nam and Davidson, 2012). The vectors are constructed such that the tags are transcribed only if and when the included genomic fragment displays positive cis-regulatory activity in developing embryo nuclei. The true efficiency of this protocol is a consequence of integrating the BAC technology with this high-throughput screening method. In practice, given the relatively limited size of the gene regions (from the genes immediately up- and downstream of the gene of interest; < 60 kb of genomic sequence) and tiling arrangement (~2 kb genomic DNA fragments with ~1 kb overlap, yielding 2x coverage), this approach permits screening intergenic and intronic sequences from two to three marker BACs per injection episode.

Results from these experiments, including the gene organization, short-fragment tiling array, and NanoString transcript quantification data for each fragment are shown in Fig.3 (for the gene *Colp3a*) and Fig.S1-7 (for all other genes in this study; Table 1). Gene *Colp3a*, which is encoded on Scaffold559, is flanked by transcripts X and Y at distances of Z and A, respectively. From this total gene region, 24 fragments were generated (nine intergenic and 15 located within the introns larger than 1.5 kb) and tested. Embryos transgenic for any of 23 of these fragments exhibited ratios of transcribed to genomically incorporated tag sequence from 1.6 - 3.0. The obvious exception is fragment #306, in which the level of transcripts containing the tag was 14 times greater than the number of tags incorporated into the genomic DNA. This fragment, which was located 6 kb upstream of translation start site for *Colp3a*, was thus defined as a “hot piece” (HP) that demonstrated significant positive regulatory function. In cases where more than one fragment showed expression copy numbers per construct that were higher than background, a microscopic observation of spatial expression was made. In all of these cases a single spatially correct high expressing fragment could be chosen.

From each of the skeletogenically active marker BACs, we recovered one or more fragments that actively promoted transcription of the GFP reporter (Fig.S1-7). However, several of these displayed insufficiently dramatic activity to invite further consideration. Due to less than complete fragment coverage of the marker BACs (likely due to genomic sequence polymorphisms), we cannot eliminate the possibility that the absence of a hot piece for a given effector gene is indicative of more distant location of cis-regulatory control sites. However, given the rarity of this outcome, distal regulatory control is likely an uncommon organizational feature in this set of sea urchin genes.

The protocol summarized in Q2 (Fig.1, and text) thus affords very clear identification of positive cis-regulatory activity in a high throughput experimental context. The co-injected and co-analyzed 129 vectors of each Nanostring screening run accommodated direct examination of >100 kb of non-coding sequence, of which each element was tested in two adjacent configurations due to the overlapping tiling array. The modular rather than point organization of the regulatory elements is attested by the observation that just as in the HP of Fig.3, neither of the adjacent 2 kb fragments surrounding a given HP usually displayed activity. Although additional fragments may show expression (see intron fragment #i01B in Sp-Hypp2998; Fig.S5), our object was not to isolate and study every cis-regulatory module servicing each gene but rather to determine a key input driving accurate expression of at least one authenticated and usually required skeletogenic cis-regulatory module for each gene. Furthermore, some fragments were eliminated because they produced high ratios of transcribed to incorporated sequences but resulted in ectopic GFP expression. The results of these screens indicated that frequently, when tested in these short tag constructs, more than one positively acting cis-regulatory module per gene displayed the capacity to drive transcription (Fig.S1-7). This study of course leaves open the possibility that other modules altogether motivate expression of the same effector genes in juvenile and adult skeletogenesis (Gao et al., 2015).

Clusters of predicted target sites are deleted to assess function.

Given this set of HP sequences that drive expression of a set of skeletogenic effector genes, the remainder of the strategy (Fig.1) was directed at identifying functional transcription factor target sites. If binding sites for given transcription factors were shown to be required, the gene(s) encoding these factors were considered to provide “inputs” into the effector gene which the HP sequences control. Our intent was not to inclusively catalogue all such inputs per gene, but rather to identify known components of the regulatory state that are directly required (or individually not required) for cis-regulatory activation of each of the eight genes in this study. To identify the highest priority sites for investigation, we employed superimposed three successive “fuzzy” discriminators (*i.e.*, none was taken as individually decisive, but, in sequence, each was used to prioritize sets of target sites to be mutated). First, HP sequences were scanned for putative target sites of known transcription factors that constitute the skeletogenic GRN regulatory state. Second, based on clustering of identified target sites, unique local occurrence in the HP sequence, or seeming peculiarities in their arrangement such as equidistant spacing or immediately proximal combinations, deletions (typically

a few hundred bp long) were made and functionally tested, either in context of the whole BAC or in context of the short HP vectors. Third, we utilized the Echinobase genome browser to co-localize these deletion results, the identified target site distributions, and ATACseq data for *S. purpuratus* embryos (http://www.echinobase.org/Echinobase/jbrowse/index.html?data=data_atac_sp). The ATACseq data (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95651) was analyzed from embryos at 18 hours-post fertilization (hpf; when most of the genes in this study are not yet expressed), and compared to embryos at 24, 30, and 39 hpf, at which times these genes are expressed only in skeletogenic cells (Fig. 2). Even though the gene may have been differentially expressed in skeletogenic cells at 24 hpf, the peak expression might be at a later time. Thus, we looked for any developmentally modulated ATACseq signals that corresponded to target sites within the HP regions that had been shown to be essential in the deletion experiments. Since the ATAC analysis was made on embryos of a different genotype that the reference one in this highly polymorphic species, we did not consider ATAC marks as absolute but rather used them in combination with the putative binding sites within the fragments to identify shorter targets. We also mutated sites that lacked ATACseq signals if others for the same factor did display such signals. An illustrative JBrowse example for *Colp3a* is shown in Fig.3b and examples for the other genes in Fig.S1-7.

Provisional target site maps are computationally predicted.

Functional inputs that control the activation of effector gene batteries genes almost always bind to multiple cis-regulatory target sites (reviewed in Peter and Davidson, 2015). Therefore, we were not concerned at missing rare site variants, but rather wished to discriminate as well as possible target sites for a given factor from similar sites of other factors. The well-characterized skeletogenic GRN defines the limited set of transcription factors expressed in this differentiating lineage that may serve as inputs to the effector genes. To identify putative binding sites for these factors, we used the binding site compilation database CIS-BP (Weirauch, et al. 2014) in combination with previously identified binding sites in sea urchins. Because the regions of interest were limited in size (we focused here only on the 2 kb hot pieces for each gene), we focused only upon the strongest position weight matrix (PWM) signals and did not use statistical applications of PWM distributions, over-represented sequence estimates or alignment matches to consensus sites. The search strings for each transcription factor are shown in Table 2. Briefly, the numerical mouse or human PWM data for given factors in CIS-BP were utilized to map the HP sequences using a sliding window. Each string of bases was analyzed using the PWM and categorized into one of three options: likely, in which the probability values based on the PWM were ≥ 0.85 (these were typically the "core sequences"); unlikely, in which any base occurred at < 0.1 frequency within the PWM; or indeterminate (all other variants).

Of the eleven (?) potential regulatory factors in the skeletogenic GRN, four were eliminated from consideration at this step. These include Crsnp2, a transcription factor for which the target site is too degenerate (Weirauch, et al. 2014) as well as Hex (a homeodomain factor) and Tel (an ETS family factor), which also exhibit binding site degeneracy and are expressed in non-skeletogenic cells. We also removed

the T-box factor Tbrain from our analysis because, although its expression is strictly skeletogenic in *S. purpuratus* (Wahl, et al., 2009), this factor was co-opted into skeletogenic function more shallowly in evolution (Minemura, et al., 2009; Gao and Davidson, 2008; Erkenbrack and Davidson, 2015). Finally, the predicted binding sites of Deadringer (Dri) were also insufficiently discriminatory. Consequently, for this we used previously authenticated sites (Amore and Davidson, 2006). Thus, we identified putative binding sites for the remaining transcription factors that mediate the terminal region of the skeletogenic GRN: Alx1, Ets1/2, FoxB, Mitf, and Tgif. The disposition of the predicted sites in the primary HP DNA sequences is shown in Fig.3, 4 and Fig.S1-7.

Deletion assays demonstrate binding site necessity.

Target site maps were constructed for the HP sequence of each effector gene (Fig.4). Given the ease with which deletions in the marker BACs can be made, we designed at least two deletions per HP sequence (indicated by the blue bars in Fig.3, Fig.S1-7). Our intent here was to further narrow the input search space to a few hundred bp by functional test. Deletions were guided by apparent irregularities in the site clustering, such as hetero- or homotypic sites, or unique site incidence. Marker BACs with specific deletions were injected into fertilized eggs; subsequent reporter expression was assayed by visualizing fluorescence in skeletogenic cells of developing embryos. In sea urchin eggs, exogenous DNA is incorporated into the host genome during cell division, which resulting in mosaic expression patterns (e.g., if DNA is incorporated during 1st cleavage, the skeletogenic expression frequency is 50% of cells; at 2nd cleavage 25%, etc.; Livant, et al., 1991). Consequently, the most robust approach to identify necessary cis-regulatory modules is not to identify quantitative decreases in transcription, but rather complete abrogation of expression. The deleted regions were considered necessary if and only if their absence in the marker BAC eliminated GFP expression (>100 embryos per experiment).

Of the eight genes in this study, results from these deletion experiments fell into two categories (Fig.4, Fig.S1-7). For five of the eight genes (*Colp3a*, *Arghap28*, *Astacin*, *Hypp2998*, and *Mitf*), deleting specific regions within the HP in the context of the marker BAC abolished GFP expression in embryos (Fig.4). This indicates that the HP sequences identified for these genes included cis-regulatory modules required for effector gene expression and that these sites were included in the deleted regions. It should be noted that these cis-regulatory modules may not represent the only functional regulatory inputs. In contrast, deletion of regions from individual HP sequences did not abolish reporter expression for the remaining three genes (*Csrnp2*, *Dri,* and *p58a*). This is consistent with the observation that, the initial HP screens for these genes revealed more than a single active cis-regulatory module, which is typical for sea urchin embryos (Lee, et al., 2007). For these cases, deletions were analyzed in the context of the individual HP fragments, and the non-deleted HP constructs served as the control, rather than the complete marker BAC (Fig.S3, S4 and S7).

Specific binding sites are mutated to assess functionality *in vivo*.

We have now efficiently identified short (a few hundred bp) regions within much larger marker BAC sequences (>100 kb) and are now the endgame of cis-regulatory analysis (see Q5; Fig.1). Three types of predictive evidence were applied to mutation design: 1) CisBP site mapping which determines what sequence and how to mutate it; 2) functional deletion evidence which determines the sequence range where mutations ought to be successful; and 3) ATACseq data which identifies the sequence range where deletions did not affect expression but where open chromatin indicates we may have missed something.

The genome browser maps containing superimposed HP sequences, ATACseq peaks, predicted target sites of interest and deletion results (Fig.3 and Fig.S1-7) were now applied to design sets of site-specific mutations to test *in vivo* and identify cis-regulatory inputs. Our objective in the preceding experiments was high-throughput, simultaneous analysis of many cis-regulatory systems; consequently, no individual dataset could be regarded as completely inclusive and we took seriously only obviously positive results. Thus, developmentally interesting local ATACseq peaks drew our attention where they coincided with identified target sites within functionally determined deletion sequences, but lack of an ATACseq peak was not considered decisive. Furthermore, at this point, again because of the throughput available, we explored all species of target site identified as above within each functional deletion element (usually 4 to 5 per HP, plus occasional potentially interesting clusters of heterologous sites. Mutated HPs were rapidly constructed by assembly of unchanged PCR fragments with one to several hundred bp long synthetic oligonucleotides containing multiple site mutations and incorporated in tag vectors (“13 tag system”, (Nam, et al., 2010). The activity of all 50 constructs generated for the eight gene set could be robustly quantitated in a modest number of injection/QPCR experiments.

Reporter construct activity was assessed by fluorescent microscopy and qPCR measurement using the tags. In each case, GFP transcripts were present at less than 1/3 of the control hot piece (Table 4). Individual or multiple binding site mutations resulted in the absence of GFP expression in embryos. But the only cluster of several different binding sites to exhibit 0% GFP expression and a significant knockdown of expression was p58A_5021-H1 which contains sites for Ets1/2, Alx1, FoxB and Mitf (Fig.S7). Mutations of all the individual transcription factor binding sites but FoxB in this fragment did not abolish expression, implying that the important input is FoxB.

Assembling a gene regulatory network.

For each of the eight genes in our case study, we have identified one transcription factor input with high confidence and followed up on it with additional experiments. The binding sites for the positive input are located in a sequence fragment that controls a high level of expression in a GFP transgene. When a portion of this fragment purported to be active based on a number of sequence features is deleted the expression is lost. And lastly the specific mutation of putative transcription factor binding sites results in quantitative loss of expression. With these data in hand we can elaborate a high confidence GRN for the 8

genes showing their individual inputs (Fig.5). At the same time, we can exclude active input of the tested ineffective mutations at least in the tested fragments.

Conclusions

We have documented, through a process of elimination an active input to each of eight effector genes in the PMC GRN. Utilizing efficient batch protocols, we have added links to the network at the highest level of confidence: mutation of documented binding sites for transcription factors known to participate in the network. This strategy is not meant to be exhaustive since other functions besides the functionally necessary site are likely to be present in the region of the studied gene sequence. These are inferred by the partial reduction in expression revealed through the mutation of other binding sites. More complete analysis of individual gene regulatory functions in sea urchin embryonic gene networks show that ancillary functions can elegantly adjust timing, spatial control and switching functions (Damle and Davidson, 2011; Wahl, et al., 2009; Yuh, et al., 2001).

Of the eight genes analyzed in this study, three are structural components of the skeletal element, three are transcription factors, one is a regulatory protein (Arhgap28) and one is a peptidase. These were originally chosen because they were among the most highly enriched in a differential expression measurement of skeletogenic cells (Barsi, 2014) and a BAC clone was readily available. It is interesting that among the most highly expressed genes would be of such a variety of functions. At first approximation it is expected that the large number of structural proteins necessary to construct a calcium carbonate spicule would dominate this class (Benson, et al., 1986; Mann et al., 2008). The diversity of up-regulated functions points to the complexity of skeletogenesis in terms of cellular activities required.

The sea urchin homolog of Mitf or microphthalmia-associated transcription factor had not previously placed in the skeletogenic GRN. In mammals, MITF is one of a family of proteins, which include TFE3, TFEB, and TFEC. There appears to be only the one form in S. purpuratus. The mouse mutant exhibits defects in pigmentation of the eyes and coat; mast cell differentiation and skeletogenic defects(Simões-Costa, 2015; Teitelbaum and Ross, 2003; Garraway et al, 2005). The cells responsible for the defects are all migratory cell types: mast cells, neural crest derived pigment cells and osteoclasts. Perhaps Mitf is part of a conserved kernel regulating proliferation in migratory cells. Active inputs to Mitf include Sox10, Pax3 and Lef1 (Boundurand et al, 2000; Yasamoto et al, 2002). Homologs of the mammalian genes are expressed at the appropriate time in the purple sea urchin to influence Mitf expression.

The distribution of active inputs to the eight gene regulatory functions described here are not organized in a strictly hierarchical fashion. Transcription factors immediately downstream of the double negative gate (Oliveri et al., 2001) are direct and necessary inputs to the eight genes as are ones downstream of the immediate ones. The role of those factors whose inhibition did not abolish expression also contribute to this complicated regulatory state. Although our analysis is too simple to fully evaluate these roles. This example

emphasizes the importance of the studying transcription factor binding to exclude indirect effects resulting from perturbations further upstream in the GRN (Davidson, 2010).

Effector genes encoding cell biology and differentiation genes do all the functional work of the cell, and together constitute the large majority of protein coding genes. Though the cis-regulatory control systems of several batteries of differentiation genes have been intensely studied, the generalities of immediately upstream effector gene control circuitry remain sketchy. The challenge of understanding effector gene control systems demands new approaches to cis-regulatory analysis, which can determine necessity as well as sufficiency, and can be multiplexed to increase throughput. Here we exploit BAC recombineering for cis-regulatory analysis, together with other recent inventions of our laboratory, to determine the regulatory inputs of a set of skeletogenic effector genes of the sea urchin embryo. The assembly and application of this new combined approach to cis-regulatory analysis reveals GRN wiring immediately upstream of the effector genes themselves.

ACKNOWLEDGEMENTS

This research was supported by NIH (P40OD010959, P41HD071837, HD037105) and the Beckman Institute through the Center for Computational Regulatory Genomics. We thank Drs. Ellen Rothenberg and Kathryn Buckley for thoughtful comments on the manuscript.

Authors' contributions:

LW, KK, RAC, AC, PD executed the experiments; PK and EHD designed method used for sequence analysis; RAC, LW, and EHD designed and interpreted the experiments; EHD and RAC wrote the paper.

The authors declare no conflicts of interest.

REFERENCES

Amore, G., Davidson, E.H. 2006. cis-Regulatory control of cyclophilin, a member of the ETS-DRI skeletogenic gene battery in the sea urchin embryo. Dev Biol 293, 555-564.

Barsi, J.C., Tu, Q., Davidson, E.H. 2014. General approach for in vivo recovery of cell type-specific effector gene sets. Genome Res 24, 860-868.

Benson, S. C., Benson, N. C., and Wilt, F. 1986 . The organic matrix of the skeletal spicule of sea urchin embryos. J. Cell Biol. 102, 1878-1886.

Bondurand, N., Pingault, V., Goerich, D. E., Lemort, N., Sock, E., Le Caignec, C., Wegner, M., Goossens, M. 2000. Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome. Human Molecular Genetics, 9: 1907–1917.

Buckley, K. M., Dong, P., Cameron, R. A., Rast, J. P. 2017. Bacterial artificial chromosomes as recombinant reporter constructs to investigate gene expression and regulation in echinoderms. Briefings in Functional Genomics, elx031, https://doi.org/10.1093/bfgp/elx031

Cameron R.A., Oliveri, P., Wyllie, J., Davidson, E.H. 2004. cis-Regulatory activity of randomly chosen genomic fragments from the sea urchin. Gene Expr Patterns 42, 205-213.

Cameron R.A., Samanta, M., Yuan, A., He, D., Davidson, E.H. 2009. SpBase, the sea urchin genome database and web site. Nucleic Acids Res 37(Database issue,D750-754). doi,10.1093/nar/gkn887

Cameron, R.A., Mahairas, G., Rast, J.P., Martinez, P. et al. 2000 . A sea urchin genome project, sequence scan, virtual map, and additional resources Proc Natl Acad Sci Unit States Am. 97, 9514-9518 .

Chen, C.G., Fabri, L.J., Wilson, M.J., Panousis, C. 2014. One-step zero-background IgG reformatting of phage-displayed antibody fragments enabling rapid and high-throughput lead identification. Nucleic Acids Res 42, e26.

Damle, S., Davidson, E.H. 2011. Precise cis-regulatory control of spatial and temporal expression of the, Alx-1 gene in the skeletogenic lineage of S. purpuratus . Dev Biol 357, 505-517.

Davidson, E.H. 2006. The Regulatory Genome. Gene Regulatory Networks in Development and Evolution (Academic Press/Elsevier, San Diego, CA.)

Davidson, E.H. 2010. Emerging properties of animal gene regulatory networks. Nature 468, 911-920.

Dixit et al., 2016, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell 167, 1853–1866.

Erkenbrack, E.M. Davidson, E.H. 2015. Evolutionary rewiring of gene regulatory network linkages at divergence of the echinoid subclasses. Proc Natl Acad Sci Unit States Am 112. E4075-4084.

Frengen, E., Weichenhan, D., Zhao B., et al. 1999. A modular, positive selection bacterial artificial chromosome vector with multiple cloning sites. Genomics 58, 250–253

Gao, F., Davidson, E.H. 2008. Transfer of a large gene regulatory apparatus to a new developmental address in echinoid evolution. Proc Natl Acad Sci Unit States Am 105, 6091-6096.

Gao, F., et al. 2015. Juvenile skeletogenesis in anciently diverged sea urchin clades. Dev Biol 400, 148-158.

Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy,S., Beroukhim, R., Milner, D. A., Granter, S. R., Du, J., Lee, C., Wagner, S. N., Li, C., Golub, T. R., Rimm, D. L., Meyerson, M. L., David E. Fisher, D. E. and Sellers, W. R. 2005. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 436: 117-122.

Geiss, G.K., et al. 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 26, 317-325.

George, N.C., Killian, C.E., Wilt, F.H. 1991. Characterization and expression of a gene encoding a 30.6-kDa Strongylocentrotus purpuratus spicule matrix protein. Dev Biol 1472, 334-342.

Gibson, D.G., et al. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343-345.

Hobert, O. 2002. PCR Fusion-Based Approach to Create Reporter Gene Constructs for Expression Analysis in Transgenic C. elegans . BioTechniques 32, 728-730.

Hollenback, S.M., Lyman, S., Cheng, J. 2011. Recombineering-based procedure for creating BAC transgene constructs for animals and cell lines. Curr Protoc Mol Biol Chapter 23,Unit 23,14.

Holmes, S., Lyman, S., Hsu, J.K., Cheng, J. 2015. Making BAC transgene constructs with lambda-red recombineering system for transgenic animals or cell lines. Methods Mol Biol 1227, 71-98.

Katoh-Fukui, Y., et al. 1991. The corrected structure of the SM50 spicule matrix protein of Strongylocentrotus purpuratus. Dev Biol 145, 201-202.

Kurokawa, D., et al. 1999. HpEts, an ets-related transcription factor implicated in primary mesenchyme cell differentiation in the sea urchin embryo. Mech Dev 80, 41-52.

Lee, P.Y., Nam, J., Davidson, E.H. 2007. Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. Dev Biol 3072, 434-445.

Livant, D.L., Hough-Evans, B.R., Moore, J.G., Britten, R.J., Davidson, E.H. 1991. Differential stability of expression of similarly specified endogenous and exogenous genes in the sea urchin embryo. Development 1132, 385-398.

Livingston, B.T., et al. 2006. A genome-wide analysis of biomineralization-related proteins in the sea urchin Strongylocentrotus purpuratus. Dev Biol 300, 335-348.

Longabaugh, W.J.R. 2012. BioTapestry, A Tool to Visualize the Dynamic Properties of Gene Regulatory Networks. Methods Mol Biol. 786, 359-94.

Lyons D., Martik, M,, Saunders, L., McClay, D. 2014. Specification to biomineralization, following a single cell type as it constructs a skeleton Integr Comp Biol 54, 723-733.

Mann, K., Poustka, A. J., Mann, M. 2008. The sea urchin (Strongylocentrotus purpuratus) test and spine proteomes. Proteome Science 6:22

Minemura, K., Yamaguch,i M., Minokawa, T. 2009. Evolutionary modification of T-brain: tbr expression patterns in sand dollar. Gene Expr Patterns 9, 468-474.

Nam, J. Davidson, E.H. 2012. Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. PLoS One 7, e35934.

Nam, J., Dong, P., Tarpine, R., Istrail, S., Davidson, E.H. 2010. Functional cis-regulatory genomics for systems biology. Proc Natl Acad Sci Unit States Am 107, 3930-3935.

Oliveri, P., Tu, Q., Davidson, E.H. 2008. Global regulatory logic for specification of an embryonic cell lineage. Proc Natl Acad Sci Unit States Am 105, 5955-5962.

Paquette, S.M., Leinonen, K., Longabaugh, W.J.R. 2016. BioTapestry now provides a web application and improved drawing and layout tools [version 1; referees, 3 approved]. F1000Research 5, 39. 10.12688/f1000research.7620.1

Peter, I.S. Davidson, E.H. 2015. Genomic Control Process, Development and Evolution (Academic Press, Elsevier, Oxford).

Peter, I.S., Faure, E., Davidson, E.H. 2012. Feature Article: Predictive computation of genomic logic processing functions in embryonic development. Proc Natl Acad Sci Unit States Am 109, 16434-16442.

Rafiq, K., Cheers, M.S., Ettensohn, C.A. 2012. The genomic regulatory control of skeletal morphogenesis in the sea urchin. Development 139, 579-590.

Sharan, S.K., Thomason, L.C., Kuznetsov, S.G., Court, D.L. 2009. Recombineering, a homologous recombination-based method of genetic engineering. Nat Protocol 42, 206-223.

Spitz, F. and Furlong, E.E.M. 2012. Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics 13: 613-626

Sucov, H.M., Hough-Evans, B.R., Franks, R.R., Britten, R.J., Davidson E.H. 1988. A regulatory domain that directs lineage-specific expression of a skeletal matrix protein gene in the sea urchin embryo. Gene Dev 20, 1238-1250.

Swanson, C.I., Schwimmer, D.B., Barolo, S. 2011. Rapid evolutionary rewiring of a structurally constrained eye enhancer. Curr Biol 21, 1186-1196.

Teitelbaum, S. L. and Ross. F. P. 2003. Genetic Regulation Of Osteoclast Development And Function. Nature Reviews Genetics, 4: 638-649.

Tu, Q., Brown, C.T., Davidson, E.H., Oliveri, P, 2006, Sea urchin Forkhead gene family, phylogeny and embryonic expression. Dev Biol 300, 49-62.

Tu, Q., Cameron, R.A., Davidson E.H. 2014. Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. Dev Biol 385, 160-167.

Wahl, M.E., Hahn, J., Gora, K., Davidson, E.H., Oliveri, P. 2009. The cis-regulatory system of the tbrain gene, Alternative use of multiple modules to promote skeletogenic expression in the sea urchin embryo. Dev Biol 335, 428-441.

Weirauch, M.T., et al. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158, 1431-1443.

Xie et al. 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Molecular Cell 66, 285–299

Yasumoto, K., Takeda, K., Saito, H., Watanabe, K., Takahashi, K., and Shibahara, S. 2002. Microphthalmia-associated transcription factor inter acts with LEF-1, a mediator of Wnt signaling.

Yuh, C.H., Bolouri, H., Davidson, E.H. 2001 . Cis-regulatory logic in the endo16 gene, switching from a specification to a differentiation mode of control . Development 128, 617-629.

Zhu, X., et al. 2001 A large-scale analysis of mRNAs expressed by primary mesenchyme cells of the sea urchin embryo. Development 128, 2615-2627.

Fig.1. Sequence of analytical procedures. Four or five successive sets of functional observation were made on the regulatory system of each gene. This was done either by microscopic observation of spatial expression of the GFP marker in individual embryos, in which the fraction of embryos examined was also recorded (S); or by quantitation of expression per construct using the multiplexed tag system in co-injected batches of embryos, measured by Nanostring as described in text (Q). At the first step, S1, efficacy of recombineered marker BACs was established. At the second, Q2, the whole intronic and intergenic region of each gene was segmented, scanned for activity, and one or more 2kb fragments displaying

transcriptional activity (“hot pieces”) were identified. At the third, S3, spatial expression of marker BACs bearing specific deletions made within the putative regulatory regions (hot piece, marked as red rectangle) were assessed for spatial expression and fraction of embryos expressing, compared to controls. In most cases localized deletions were found which abolished marker BAC activity, and the analysis of these genes proceeded to step Q5. At the fourth step, S4, in cases where individual deletions failed to abolish BAC expression, these deletions were tested for spatial expression in the context of the individual active 2kb fragments. In the final step Q5, site clusters revealed to be functional in steps S3 or S4 were subjected en masse to site specific mutational analysis.

Fig.2. Illustrations of in vivo expression of recombineered marker BACs carrying GFP coding sequences inserted immediately following ATG in the genes included in this study. BACs were injected into fertilized eggs at about 500 copies, and in each case the cells expressing them are skeletogenic. At mesenchyme blastula stage these cells form bilateral clusters on the floor of the blastocoel; as gastrulation proceeds they form characteristic rings surrounding the archenteron; later form syncytial chains as they secrete skeletal rods within which GFP diffuses from cell to cell. Gene names are shown on figure panels. Genomic maps of BACs and gene regions can be accessed by reference to BAC designations as listed in Table 1 in Echinobase (http://www.echinobase.org/Echinobase/).

Fig.3. Example of high throughput tag vector screen for cis-regulatory activity in a marker BAC. As described in text marker BACs, here containing the Colp3a gene, were segmented into ~2kb fragments forming an overlapping tiling array (black rectangles). Large exonic regions were excluded. Gene structure (exons) are represented by yellow boxes on the black (transcriptome) line. Flanking the black (transcriptome) line is the relevant region of the BAC, i.e., the gene plus the flanking upstream and downstream regions. The flanking yellow rectangles indicate the next gene upstream and downstream in the BAC. For identities of surrounding genes and BAC maps see http://www.echinobase.org/Echinobase/.

Each 2kb fragment shown as a vertical bar was incorporated in a tag vector. The mixture of all 2kb-Tag vectors from 8 genes were injected into fertilized eggs. Transcribed tag DNA and genomically incorporated tag DNA for all vectors were measured simultaneously by Nanostring. The ratio of these values, which indicates the transcriptional activity driven by the sequence in each respective vector normalized to the amount of vector DNA incorporated, are shown at right in the histogram (Fig.3A). 2kb fragments that are thus revealed to include active cis-regulatory modules are shown in red.

Diagrams redrawn from JBrowse screen captures showing various features mapped to the reference genome sequence (Fig.3B). The BAC is shown in green. The red line depicts the DNA sequence of the active fragment “hot piece” in the BAC. Deleted regions are shown in blue below it. Symbols identifying the deletions are the ones used in the test and tables. Vertical bars represent the position and number of ATACseq reads mapped. The position of the modified CIS-BP transcription factor binding sites are indicated in light green. Mutated sites are indicated in orange.

Fig.4. Deletion and Mutation Mappings. Schematic diagrams of the individual BAC clones bearing the genes under study. The scale is approximate. Deleted regions are shown in the blue line under the black line depicting the hot piece sequence. Symbols identifying the deletions are used in the test and tables.

BAC deletions are shown as upper case and hot piece deletions are lower case. Vertical bars represent the position of the transcription factor binding sites tested in this study. Erg binding sites in most cases overlapping with Ets1/2 ones. Thus we simultaneously mutated both Ets1/2 and Erg sites in mutation analysis. Similar to the case for Csrnp2 sites, the sites for Dri were too ambiguous to be mapped For clearer illustrations, we omitted Erg and Dri sites.

Fig.5. The PMC gene regulatory network expanded in this study. This GRN was produced in the BioTapestry Editor. Many of the upstream genes are removed for simplicity. The present diagram starts at the double-negative gate (Pmar1-Hesc). The genes in this study (Colp3a, Arhgap28, Astacin1, Csrnp2, Dri, Hypp_2998, Mitf, P58a) are along the bottom with other known components removed except for Sm50 to anchor the downstream level. One of the studied genes, Mitf encodes a transcription factor and is placed in the middle control region. The possible inputs to the studied genes that were shown to reduce but not eliminate expression by means of the mutation experiments are aligned below each studied gene symbol.

Table 1. Genes included in this study

Gene name	ID[1]	Function of encoded protein
Colp3a	SPU_003768	Collagen type 4, forms basement membrane cellular mesh
Arhgap28	SPU_027628	GTP activating protein, RhoA-dependent actin organization in cell contractility
Astacin1	SPU_019655	Metalloprotease, plasma membrane, affects extracellular proteins/growth factors
Csrnp2	SPU_014613	Probable transcription factor
Dri	SPU_005718	Transcription factor
Hypp_2998	SPU_018407	Transmembrane organic spicule biomineral matrix protein
Mitf	SPU_008175	Transcription factor
P58a	SPU_000439	Biomineralization protein

Notes to Table 1: Gene ID number is given as “SPU” designation, from the public genome database for Strongylocentrotus purpuratus http://www.echinobase.org/Echinobase/ . The most comprehensive analyses of skeletogenesis specific proteins in this organism are to be found in refs. (Rafiq, et al., 2012; Zhu, et al. 2001; Livingston, et al. 2006). [1] ID refers to the Echinobase unique identifiers at www.echinobase.org.

Table 2. Binding sites: CIS-BP Position Weight Matrix

Gene name	Alx1	Ets1/2	FoxB	Tgif	Erg	Mitf
CIS-BP ID	M6141	M6221	M0737	M5933	M5401	M0208

Base 1	T	V	R	T	V	R
2	A	G	T	G	G	N
3	A	G	A	A	G	C
4	T	A	A	C	A	A
5	B	A	A	A	H	B
6	Y	R	Y	S	R	R
7	V	N	A	C		N
8		D				G

Notes for Table 2. IUPAC symbols for nucleotide combinations: B= C, G, T, not A; V= A, C, G, not T; D= A, T, G, not C; H= A, C, T, not G; Y= C, T, not A, G; R = A, G, not C, T; S= C, G, not A, T; N = any nucleotide.

Table 3. Expression of transcription factor binding site mutations in the positive fragment.

Hot Piece[1]	Mutations[2]	% GFP[3]
Colp3a_306	A1 (Ets1/2)	29%
	A2 (Tgif)	50%
	A3 (Mitf)	0%
	A4 (FoxB)	38%
	A5 (Cluster)	41%
Arhgap28_504	B1 (Tgif)	0%
	B2 (Mitf)	40%
	B3 (FoxB)	22%
	B4 (Alx)	20%
Astacin1_504	C1 (Dri)	7%
	C2 (Mitf)	22%
	C3 (Alx1)	24%
	C4 (FoxB)	15%
	C5 (Ets1/2)	0%
Csrnp2_i10	D1 (Cluster1)	50%
	D2 (Alx1)	50%
	D3 (Mitf)	0%
	D4 (Cluster2)	34%
Dri_507	E1 (Cluster1)	22%
	E2 (Alx1)	40%
	E3 (Ets1/2)	14%
	E4 (Mitf)	30%
	E5 (Cluster2)	22%
	E6 (FoxB)	0%
Hypp2998_306	F1 (Alx1)	0%

Mitf_506	F2 (Ets1/2)	15%
	F3 (Tgif)	19%
	G1 (Dri)	41%
	G2 (Ets1/2)	33%
	G3 (Mitf)	0%
	G4 (FoxB)	33%
p58a_5021	H1 (Cluster)	0%
	H2 (Alx1)	50%
	H3 (Ets1/2)	16%
	H4 (Tgif)	50%
	H5 (Mitf)	50%

[1] The Hot Piece determined by NanoString nCounter measurements. [2] The individual binding site mutations made in the Hot Piece. [3] Percent of embryos expressing GFP observed by microscopy.

Table 4. Functional transcription factor binding site mutations in the positive fragment.

Hot Piece[1]	Mutations[2]	% GFP[3]	% KO[4]	Note
Colp3a_306	A3 (Mitf)	0%	84%	5 Mitf, 2 in and 3 outside of deletion 2.
Arhgap28_504	B1 (Tgif)	0%	70%	1 Tgif in deletion2
Astacin1_504	C5 (Ets)	0%	89%	4 Ets1/2, outside of deletion 2
Csrnp2_i10	D3 (Mitf)	0%	88%	5 Mitf, 2 in and 3 outside of deletion2
Dri_507	E6 (FoxB)	0%	63%	1 FoxB in deletion3
Hypp2998_306	F1 (Alx1)	0%	75%	8 Alx1 in deletion2
Mitf_506	G3 (Mitf)	0%	84%	8 Mitf in deletion1
P58_5021	H1 (Cluster)	0%	92%	Group1: 3 Alx1, 2 Ets1/2 Group 2: 4 Ets1/2, 2 Mitf, 2 Tgif, 1 FoxB

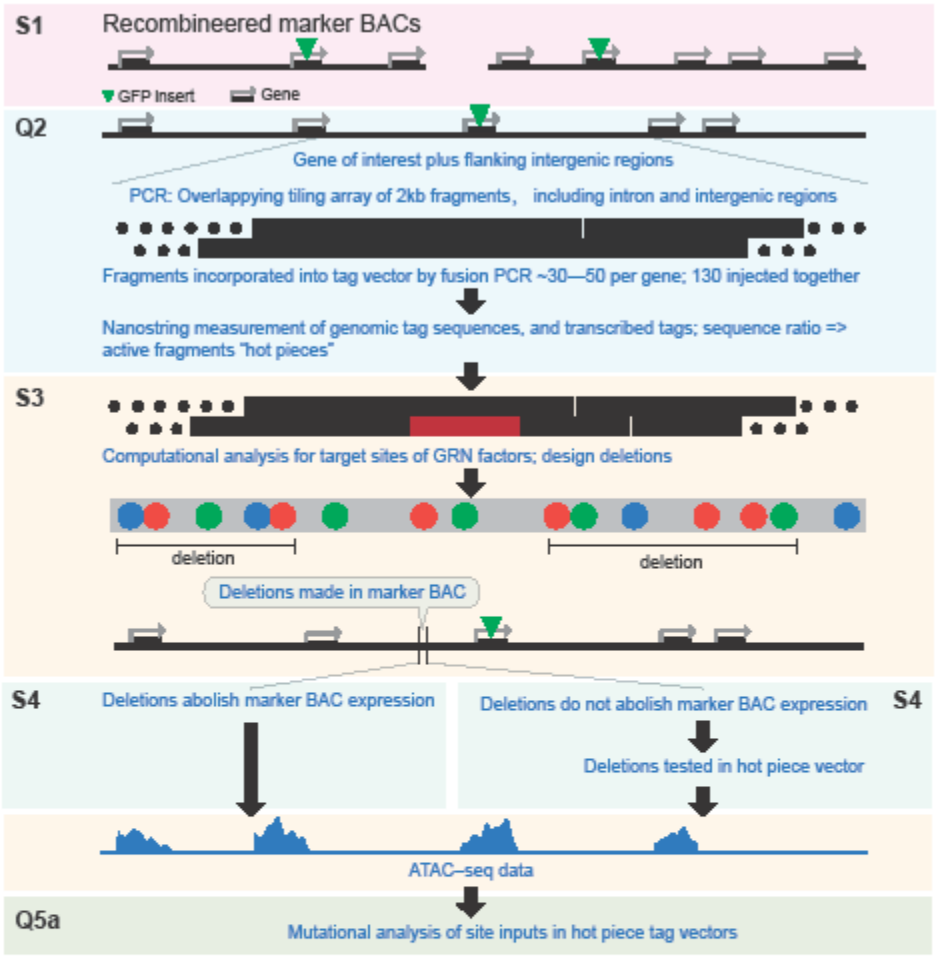
[1] The Hot Piece determined by NanoString nCounter measurements.

[2] The individual binding site mutations made in the Hot Piece.

[3] Percent of embryos expressing GFP observed by microscopy.

[4] The percent knockdown of GFP expression measured by qPCR. See Supplementary Table 1 for genomic coordinates of these fragments.

Highlights:
Recombineered BAC clones are useful for rapid gene regulatory network analysis.
High throughput methods using BACs quickly reveal cis-regulatory function
BAC reporters can assess the necessity or sufficiency of cis-regulatory modules.



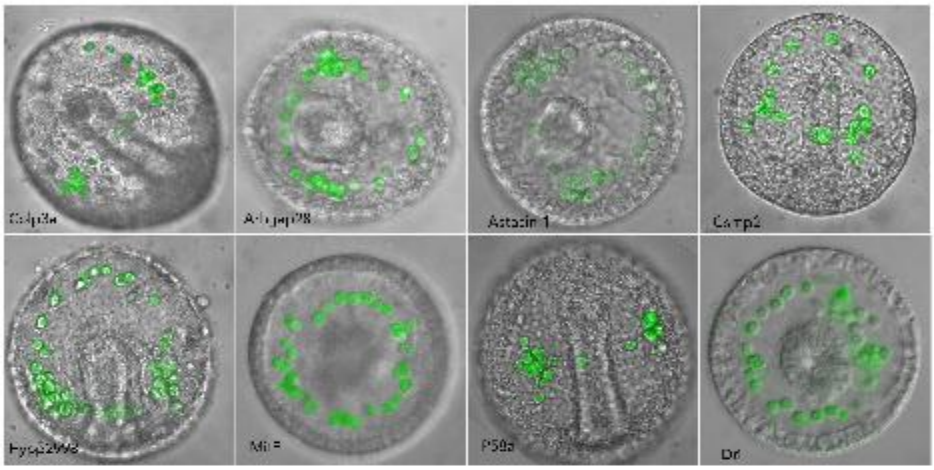


Figure 3A.

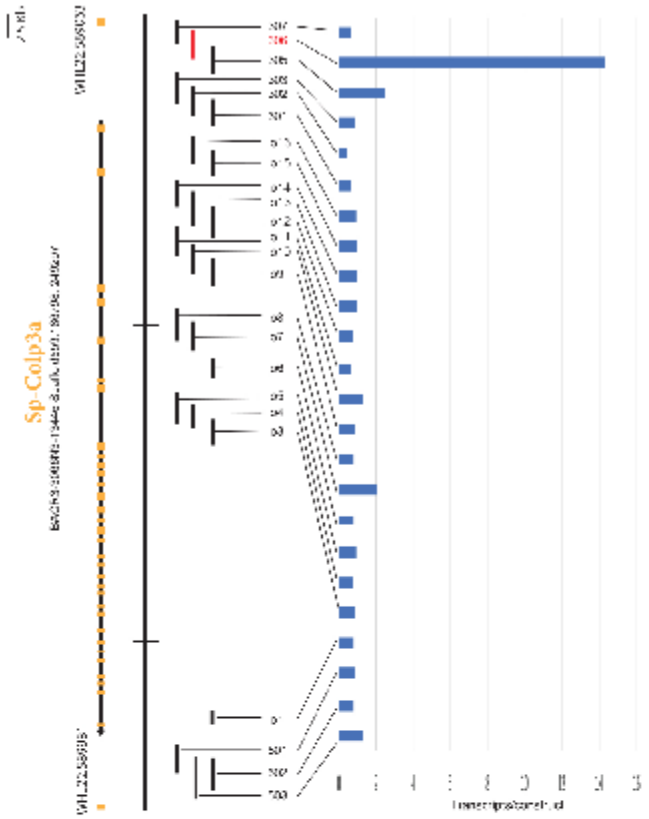


Figure 3B.